

Simranjeet Singh

LinkedIn/simranjeet-singh
Leetcode/loxa_
Codeforces/loxa

Email: sjsingh5360@gmail.com
Mobile: +91-8371909111

SUMMARY

AI Engineer with 4 years building production-grade LLM, RAG, and multi-agent AI systems; experience in scalable retrieval pipelines, vector databases, and enterprise GenAI infrastructure supporting 50+ engineering teams.

EDUCATION

- Indian Institute of Technology Kharagpur** Kharagpur, India
Dual Degree (5 Yrs.) – Electronics and Electrical Communication Engg.; GPA: 8.40 2017 – 2022
- Sri Aurobindo International School** Patiala, India
Higher Secondary Education; Percentage: 92% 2017

EXPERIENCE

- Qualcomm** Hyderabad
Engineer July 2022 – Present
 - Code Review Assistant:**
 - * Automated PR review across GitHub, Gerrit, and Perforce using cloud LLMs (**Claude, GPT, Gemini**) — surfacing logic bugs, security issues, and style violations. **Onboarded 50 teams across 300+ projects, saving 30–45 mins per review cycle.**
 - * **Non-Agentic:** Diff-aware pipeline optimized for latency with context compression (prioritizing diffs, trimming low-signal content) to stay within LLM context limits; reviews generated in **1–2 min.**
 - * **Agentic:** Dual-agent architecture running in a **sandboxed environment where the agents have direct access to the full codebase** — step-budgeted *Issue Agent* with symbol-level indexing for accurate cross-file context + *Fix Agent* validated by `pylint/clang-tidy` before posting; optimized for correctness over speed, end-to-end review in **~10 min.**
 - * **Local Review Skill:** Recently introduced a code review skill for local (pre-push) reviews, built on the open-source **Deep Agents harness** so it is portable across agent harnesses such as **Claude Code** and **Codex.**
 - * **Scalable Serving:** Deployed on **Kubernetes** with a request queue and distributed worker nodes to handle concurrent PR reviews across 300+ projects.
 - Synthetic Data Generation & LLM Fine-Tuning:**
 - * Built a modular ingestion layer (PDF, DOCX, Confluence, Jira, SharePoint) and generated **67K QA pairs** using Mistral/LLaMA with zero-shot, few-shot, and chain-of-thought prompting; quality-filtered via ROUGE scoring and LLM-as-evaluator.
 - * Fine-tuned LLaMA using **PEFT/LoRA** (HuggingFace Transformers); achieved **15% improvement** in RAGAS scores (faithfulness, answer relevance, context precision) on an **80/10/10 train-val-test split** over the synthetic dataset.
 - Agentic RAG Developer Toolkit:**
 - * Built a reusable Python package for RAG-powered agentic workflows using **Agno** for multi-agent orchestration; teams inherit base classes and implement only domain-specific logic — abstracting away vector DB lifecycle (**FAISS, Chroma**), chunking strategies, and embedding pipelines.
 - * Benchmarked retrieval configurations (vector search, TF-IDF, BM25, hybrid) and reranking models (**BGE-Reranker**); achieved **5% improvement in Top-K retrieval precision and recall** over baseline in enterprise document search workloads.
 - * Integrated **Langfuse** for LLM observability — tracing retrieval accuracy, latency, and token costs in production to identify pipeline bottlenecks and drive continuous improvements.
 - Data Mining Pipeline:**
 - * Designed dataset aggregation with custom computations on AWS, achieving **60% cost reduction** (>\$100k/month); automated schema validation boosted monthly active users by **40%.**
 - * Diagnosed and resolved dataset correctness issues caused by build version changes triggering logging schema drift; implemented field-level validation to ensure correct capture of fields across version updates, improving downstream data reliability.
- Qualcomm** Hyderabad
Software Engineering Intern May 2021 – July 2021
 - Test Framework Design:** Created high-level class diagrams for revamping the unit test framework using **Visio**, and documented detailed call flows with **MscGen** to streamline the testing process.
 - Unit Testing and Coverage:** Analyzed existing tests with **CoveragePy** to identify untested code paths; designed and implemented comprehensive **Pytest** suites raising coverage to **90%.**

SKILLS SUMMARY

- LLMs & RAG:** Agno, HuggingFace Transformers, PEFT/LoRA, RAG, Corrective RAG, Hybrid Search (BM25 + Vector), BGE-Reranker, RAGAS, FAISS, Chroma, Langfuse, Prompt Engineering, LLM Evaluation, Deep Agents
- ML / DL:** PyTorch, Scikit-Learn, Keras, NumPy, Pandas, Polars, Matplotlib
- Systems & Cloud:** Prompt Caching, Token Optimization, Scalable Retrieval, Vector Databases, Model Serving, AWS, Docker, Jenkins, Kubernetes
- Languages:** Python, C, C++, SQL
- Relevant Courses:** Deep Learning, Machine Learning, AI, NLP, Computer Vision, Probability & Statistics, Algorithms & Data Structures