

Neerav Jha

Noida, Uttar Pradesh | +91 98708 17514 | neeravjha444@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

SUMMARY

Software Engineer with production experience building LLM-orchestrated conversational AI systems and backend integrations at enterprise scale. Skilled in RAG pipelines, tool/function-calling orchestration, and resilient REST API design; shipped a platform-first ServiceNow integration improving reliability and reducing recovery time. Strong systems fundamentals, with hands-on experience diagnosing and fixing production performance bottlenecks.

WORK EXPERIENCE

Yellow.ai

Associate Software Engineer

Bangalore, India

March 2025 – Present

- Built and productionized multi-channel conversational AI systems (chat & voice) using RAG pipelines and LLM orchestration (function calling, tool routing, caching), improving response latency and containment rate.
- Designed and shipped the platform's first ServiceNow integration (OAuth 2.0, event-driven sync, resilient REST APIs), improving API reliability and throughput while reducing mean time to recovery (MTTR).
- Added performance instrumentation across orchestration services to surface latency and reliability regressions before they impacted customers.

Munsow Technologies

Python Developer Intern (Backend Engineering)

Hyderabad, India

Jun 2024 – Nov 2024

- Built backend services (Flask, MongoDB) for an AI-powered flashcard learning platform, integrating the OpenAI API for automated content generation; implemented RBAC and RESTful APIs supporting 1,000+ registered users and 150+ concurrent sessions.
- Owned API testing, debugging, and performance validation using Postman, ensuring reliable request handling and consistent data across the platform.

PROJECTS

LLM Gateway — Multi-Provider LLM Routing Gateway

Python, FastAPI, Redis, Vector DB, SQLite | [Github](#) · [Demo Video](#)

- Built and deployed an OpenAI-compatible multi-provider LLM gateway with automatic fallback, two-layer caching (exact-match + semantic via vector similarity), sliding-window rate limiting, and cost tracking; fixed a connection-pooling bottleneck that cut p50 latency from 3.8s to 473ms.
- Implemented semantic caching (embeddings + cosine similarity) with a threshold calibrated from measured data (paraphrases ~0.89–0.92 vs. unrelated ~0.56), serving cached answers for reworded prompts.

taskq — Distributed Task Queue

Python, FastAPI, Redis, SQLite | [Github](#)

- Built a distributed task queue with priority scheduling, idempotent enqueue, exponential-backoff retries with dead-letter handling, and backpressure (503 + Retry-After at queue depth cap).
- Designed job leasing + heartbeat-based failure detection so a hard-killed worker's job is automatically rescued by another worker within seconds — verified via a live kill-worker test; load-tested at ~520 enqueue req/s locally (p50 28ms).

EDUCATION

HNB Garhwal University

Bachelor of Technology, Computer Science and Engineering — CGPA: 8.0

Srinagar, Uttarakhand

Aug 2024

TECHNICAL SKILLS

Languages: Python, Java, C++, C, SQL, JavaScript

Backend Frameworks: FastAPI, Flask, Node.js, Express.js, REST API Design, OAuth 2.0

AI/LLM & Systems: RAG Pipelines, LLM Orchestration, Vector Databases, System Design, Async/Concurrent Programming

Databases, Cloud & Tools: MongoDB, MySQL, PostgreSQL, Redis, SQLite, Docker, AWS, Git, Postman

ACHIEVEMENTS

- 1,200+ DSA problems solved across LeetCode and HackerRank | 5★ Python on HackerRank

CERTIFICATIONS

Palo Alto Networks Certificate Course on Cybersecurity • Cognizant Agile Methodology Virtual Experience Program • HackerRank SQL (Intermediate) Certificate