

Varun Negi

Gurugram, India

varun.py10@gmail.com — +91-8799707841 — LinkedIn — GitHub — Portfolio

Professional Summary

AI Engineer and Backend Engineer building production LLM applications, agentic AI systems, RAG pipelines, and scalable backend services with Python, Java, FastAPI, Spring Boot, PostgreSQL, MongoDB, AWS Redshift, vector search, and microservices.

Technical Skills

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

AI / ML: LLMs, Agentic AI, AI Agents, RAG, Multimodal RAG, LangChain, LangGraph, Prompt Engineering, Vector Search, Embeddings

Backend: Spring Boot, FastAPI, Node.js, REST APIs, Microservices, Distributed Systems

Databases: PostgreSQL, MongoDB, pgvector, AWS Redshift

Cloud / DevOps: AWS, Docker, Git, LangSmith

Frontend: React.js, Next.js

Professional Experience

Stanza Living

Dec 2025 – Present

Software Development Engineer I

Gurugram, India

- Architected **Support 3.0**, a LangGraph-based agentic AI support system using intent routing, domain-specific agents, conversation memory, and dynamic tool calling to resolve resident queries before ticket creation.
- Built a hybrid RAG pipeline using pgvector, lexical reranking, and module-scoped retrieval to ground responses in support policies, SOPs, FAQs, and domain documentation.
- Developed configurable LLM orchestration with versioned prompts, model selection, multimodal image understanding, guarded ticket actions, and LangSmith tracing for debugging and quality improvement.
- Owned Java and Spring Boot microservices across 5 business domains, supporting 3,000+ daily users with zero critical downtime.

Stanza Living

Jun 2025 – Dec 2025

Software Development Engineer Intern

Gurugram, India

- Designed and deployed **DataSense**, an LLM-powered analytics agent using LangChain, FastAPI, and AWS Redshift that converts natural-language questions into SQL across 400+ tables, saving 20+ hours per week for 10+ teams.
- Improved SQL accuracy on complex multi-table queries with schema-aware RAG pipelines, contextual retrieval, and prompt templates for business stakeholders.
- Improved performance and data accuracy of a consumer website synchronization pipeline by 30% through query optimization and backend data processing improvements.

Tailor Talk

Mar 2025 – Jun 2025

Software Development Engineer Intern

Remote

- Built Python and FastAPI-based LangChain agent workflows that managed 3,000+ daily calendar conversations and contributed to 300+ monthly bookings through automated scheduling.
- Implemented rate limiting, reliability controls, and reporting automation that maintained 99%+ uptime and reduced manual data-entry work.

Education

Delhi Technological University

2021 – 2025

Bachelor of Technology in Engineering Physics, Minor in Computer Science

New Delhi, India

CGPA: 7.78/10

Awards and Achievements

- Won **Analytics Ace: Best Use of Database Intelligence** at Dine with DevOps IV, Stanza Living, for building an LLM-based analytics agent.
- LeetCode rating: 1732 (Top 10%) with 850+ problems solved; CodeChef rating: 1633 with Global Rank #157.