

Tanish Vashisth

9911584966 | tanishvashisth@gmail.com | github.com/bogusdeck | leetcode.com/bogusdeck | linkedin.com/in/tanishvashisth | Delhi, India

Summary

Backend-focused engineer with 2+ years shipping production LLM features (OpenAI API) at Instahyre, plus independent agentic LLM systems: a planner/executor game agent (MineMatrix), a local-LLM RAG CLI (Rune), and an MCP server for autonomous code review (mcp-phabricator). Strong Python/Django/FastAPI backend with Elasticsearch, Celery, and Go. Targeting GenAI/LLM and forward-deployed roles owning AI feature delivery end-to-end.

Experience

Software Development Engineer 1 (SDE-1)

Dec 2024 – June 2026

Instahyre

Noida, Delhi

- Owned the entire Jobs module end-to-end; built **Magic Boost** - a recommendation engine suggesting optimal parameters to expand candidate pools - and **Market Signal** - an LLM-powered feature using the **OpenAI API** to generate JD-aware recruiter insights from unstructured job descriptions; also improved skills and boolean matching logic.
- Shipped **Fine-Tune** - a production GenAI feature using the OpenAI API and prompt engineering to let employers refine candidate searches via natural-language signals - boosting candidate relevancy by **70%** and pool size by **40%**; built the pipeline aggregating OpenAI outputs into weekly reports.
- Scaled critical APIs from **750ms to 120ms** average latency (~84% reduction) by replacing Django ORM with optimized raw SQL queries, cutting overall API response times by ~30%.

Back End Developer

May 2024 – Nov 2024

Xtage

Gurugram

- Solely built the entire backend for **Pearch** - a Shopify collection-sorting plugin with **1,000+ installs** - featuring an ML-assisted ranking engine that surfaces the most relevant products to shoppers by combining rule-based signals (discount, timing, rating, variants, deadstock) with store-owner-configurable parameters.
- Optimized API to handle **200–300 customer visits/hour** using Celery async queues; integrated Shopify subscription billing API for recurring pricing; deployed infra on AWS/DigitalOcean via Docker and Nginx; collaborated cross-functionally with frontend, UX/UI, and PM, delivering full API documentation.

Python Intern

Feb 2023 – Apr 2023

Callisto Infotech

Remote

- Built full-stack MERN applications, including rashtriyatv.com - a live news platform with a custom admin editor - plus REST API endpoints with Django REST Framework and JWT authentication.

Projects

Fasfas | *Django, AngularJS, React Native, Celery, Elasticsearch* | fasfas.in | Play Store | App Store

- Full-stack fashion e-commerce platform (Django REST backend, AngularJS frontend, React Native app) with Elasticsearch discovery and Celery async processing; built a collaborative-filtering recommendation engine (co-purchase similarity from order history) exposed via product and personalized recommendation endpoints.

MineMatrix | *Python, Ollama, RAG, Agentic Planning, Node.js* | [GitHub](https://github.com)

- Built an autonomous LLM agent that plays Minecraft end-to-end: a local Ollama-served model (qwen2.5) plans against a hand-designed goal hierarchy while a deterministic 5-node runtime (Rule Guard, Executor, Watchdog, Completion Check, async Planner) handles safety, execution, and stall recovery on a 2-second tick loop, bridged over HTTP to a Mineflayer (Node.js) game-state/action layer.
- Designed a RAG layer retrieving Minecraft-domain knowledge into the planner's prompt context, plus a Goal Guard that stops the LLM from re-selecting completed goals or getting agent-trapped in idle states.

Rune | *Go, Ollama, RAG* | [GitHub](https://github.com)

- CLI tool in Go enhancing local LLM capabilities via Ollama - implements a RAG pipeline for semantic search over personal notes, maintains conversation context across sessions, and enables fully offline AI-assisted querying.

mcp-phabricator | *TypeScript, Model Context Protocol, LLM Tool Use* | [GitHub](https://github.com)

- Open-source MCP server enabling AI agents (Claude, Gemini, Windsurf) to autonomously review code diffs with stack-aware prompts, generate task summaries, and manage Phabricator workflows via tool calls - published on Smithery, installable via Claude Code.

Technical Skills

GenAI / LLM: OpenAI API, Ollama, RAG Pipelines, Agentic Systems, Prompt Engineering, MCP, LLM Tool Use, Gemini SDK, Vector/Semantic Search

Languages: Python, JavaScript, TypeScript, SQL, Go, C++

Frameworks & Tools: Django, FastAPI, Flask, Node.js, React.js, Next.js, AngularJS, Celery, Elasticsearch, Tailwind CSS

Infrastructure: AWS, Docker, Nginx, Kibana, PostgreSQL, MySQL, Redis, DigitalOcean

Dev Tools: Git, Postman, Neovim, Phabricator, Arcanist

Libraries: pandas, NumPy, Matplotlib

Education

Bachelor of Technology in Information Technology

2020 – 2024

Krishna Engineering College (AKTU)

CGPA: 7.54