

Varad Khonde

Backend Engineer

varadkhonde@gmail.com | Maharashtra, India | [Website](#) | [LinkedIn](#) | [GitHub](#)

SUMMARY

Backend engineer with 4 years of Python experience, the last 2+ shipping production systems at a GenAI services startup. Strong in FastAPI services, real-time APIs (SSE, WebSockets), authentication and multi-tenancy, and Azure deployments. Known for taking ownership from design through deployment, and for raising the engineering bar of the team through code review, technical mentoring and hiring.

EXPERIENCE

Backend Engineer | Newtuple Technologies

Aug 2024 – Present

SQL Chat Agent text-to-SQL agent on Azure OpenAI / Foundry

- Added **SSE streaming** of answers and tool calls to an existing agent. Responses are saved only once complete, and interrupted streams return a clear retry message.
- Distributed requests across multiple Foundry agents using **round-robin load balancing through Azure API Management**, with conversations shareable across agents.
- Cut the base system prompt from **~20k to ~7k tokens** by moving the schema (2 schemas, ~30–40 tables), metrics and business rules into a database that the agent fetches on demand through a tool call. Eval accuracy improved from **~86% to ~89–90%**.
- Built an **evaluation pipeline** on ~150 golden question and SQL pairs curated by the testing team. Responses are generated through the live API, scored by an **LLM judge** and verified by a tester after every run. Accuracy tracked from **80% to 91%**; runs on every prompt change.
- Built an admin-only **trace view** (role enforced in the backend) showing the executed SQL and a result preview, with the previewed row count capped through configuration.
- Automated prompt management: prompt files are published to Foundry agents through an Azure DevOps pipeline.

Child Care Homes Management App multi-tenant, multi-service platform for a UK client

- Designed a centralized **Auth0 service** with multi-tenancy, RBAC, organization and user-level access control and permission groups, used across all backend services.
- Built the multi-service **FastAPI** backend (controller–manager architecture, standardized API schemas, consistent exception handling), deployed with Docker and GitHub Actions to Dev and Prod on Azure VMs behind **App Gateway, a load balancer and WAF**.
- Used **Redis pub/sub** for inter-service messaging and Redis caching of authenticated sessions to avoid repeated Auth0 calls.
- Built real-time collaboration features: **edit locking**, threaded notes and events, and online-user tracking, over WebSockets.

RAG Chatbot chatbot for a US national website, on Azure AI Search

- Built a retrieval-augmented chatbot over SharePoint content, with an ingestion pipeline that **auto-indexes PDF and DOCX files into Azure AI Search**.
- Implemented **semantic, hybrid and keyword retrieval**, returning citation-backed answers.
- Built **WebSocket** services for real-time text and speech chat with bidirectional streaming and persistent sessions, using Azure OpenAI Realtime APIs, Vapi and Azure Speech services.

Technical Leadership & Team Support

- Provided senior oversight on a client workflow-engine delivery (4–5 developers): integrated the company's internal workflow-engine project into the client repository via **git subtree** and owned the deployments.
- Guide 4–5 junior developers on architecture, git and deployments. Review **~20–30 PRs a week** on my project and 3–5 on others.
- Conducted **50+ technical interviews** from freshers to senior level, using a structured, resume-driven interview framework I designed.
- Proactively spotted projects running without database backups and enabled them.
- Researching private LLM hosting on RunPod to evaluate self-hosted inference options.

Python Developer (Freelance) | Remote

Oct 2022 – Aug 2024

- Built a **YOLO training app** that lets non-technical users train and deploy object-detection models.
- Built an **OpenAI Whisper** transcription pipeline, an OCR and table-extraction tool for paper documents, and a Selenium, Flask and Redis product-comparison tool.

SKILLS

Languages	Python, SQL
Backend	FastAPI, Flask, REST APIs, WebSockets, SSE, microservices
Data	PostgreSQL, Redis, Azure AI Search
Cloud & DevOps	Azure (VMs, App Gateway, WAF, API Management, Foundry), Docker, GitHub Actions, Azure DevOps, Git
Auth	Auth0, RBAC, multi-tenancy
AI integration	Azure OpenAI, Foundry agents, LangChain, RAG, LLM evaluation

EDUCATION

B.Tech Government College of Engineering, Aurangabad	2019 – 2023
--	-------------