Sonu Kumar

Gurugram, Haryana, India
+91-9693711130 sonukumarcode@gmail.com linkedin.com/in/sonukumarpandit github.com/Sonupandit9693

SUMMARY

Senior Software Engineer and Python Backend Developer with 2+ years of experience designing scalable, AI-integrated systems for enterprise applications. Specialized in API development, LLM orchestration, and system architecture using FastAPI, Flask, and LangChain. Proven success delivering RAG pipelines, multi-agent tools, and full-stack SaaS applications in cloud-native environments. Strong in Agile/Scrum, SDLC, CI/CD pipelines, and cross-functional collaboration.

TECHNICAL SKILLS

Programming Languages: Python, JavaScript, TypeScript, SQL
Frameworks: FastAPI, Flask, Django, Celery, SQLAlchemy, Alembic
Databases: PostgreSQL, MySQL, MongoDB, Redis, MariaDB, Qdrant
Backend & APIs: REST APIs, GraphQL, gRPC, Microservices, JWT Auth, WebHooks
DevOps & Cloud: Docker, AWS (EC2, S3, Lambda), Kubernetes, Terraform, Jenkins, CI/CD
AI & LLM Tools: GPT-4, LangChain, AutoGen, Whisper API, n8n, RAG
Frontend: React.js, Next.js, Vue.js, Tailwind CSS
Tools: Git, Swagger, Postman, Docker Compose, Superset, RabbitMQ, VS Code

EXPERIENCE

Cyfuture India Pvt Ltd **July 2024 – Present**
Software Engineer – Applied AI & Backend Systems *Noida, India*

- Led the internal development of a **RAG-based enterprise knowledge platform** using **LangChain**, **GPT-4**, and **Qdrant**, enabling natural-language search across business-critical documentation with **80–90%** improved relevance.
- Designed and implemented modular FastAPI microservices architecture for document parsing, OpenAI embedding generation, and vector indexing with real-time monitoring dashboards in **Apache Superset**.
- Delivered comprehensive ERP modules for Public Sector Undertaking (PSU) clients serving 5000+ users using Frappe framework and Python, covering payroll automation, leasing management, and GST compliance workflows, reducing cross-functional processing time by **40%**.
- Engineered intelligent automation system for CRM lead capture using OpenCV and GPT-4 for visitor card text extraction, reducing manual marketing data entry by **70%** and accelerating inbound lead processing workflows.
- Optimized high-volume SQL queries and **MariaDB** database configurations, implementing performance tuning strategies that reduced dashboard query latency by **30%**.
- Developed multi-agent workflow orchestration systems for enterprise automation, integrating various AI agents for document processing, data extraction, and workflow management.

Extension Technologies Pvt Ltd **July 2023 – June 2024**
Software Developer *Delhi, India*

- Engineered scalable Flask REST APIs with multithreading and rate-limiting for real-time booking engine, successfully handling **10,000+** concurrent users with **99.9%** uptime.
- Developed real-time synchronization middleware between ERPNext and Zoho systems, automating UI rendering through dynamic metadata feeds and reducing frontend latency by **50%**.
- Engineered a scalable credit allocation engine in Flask, used by 5,000+ users, improving promotion tracking by **20%.**
- Integrated third-party APIs including WhatsApp Business API, Shopify API, and DocuSign for automated lead synchronization and approval workflows.
- Implemented secure OAuth-based authentication workflows for internal document signing processes, reducing document turnaround time by **30%**

PROJECTS

GenAI Meeting Assistan | *Next.js 15, FastAPI, PostgreSQL, Redis* **May 2025 – Ongoing**

- Architected comprehensive AI-powered meeting platform with multi-modal capabilities including transcription (Whisper API), summarization (GPT-4), and automated task generation for intelligent post-meeting workflows.
- Engineered robust FastAPI backend with JWT-secured APIs, SQLAlchemy ORM, and Alembic migrations for scalable user and meeting management.
- Implemented asynchronous audio processing pipeline using Celery, Redis, and Slack integration for automated follow-ups and task distribution.
- Deployed full-stack containerized application using Docker Compose with test coverage (85%+) via PyTest.
- **GitHub:** `meeting-ai-assistant`

Enterprise RAG Search Platform (Internal) | *LangChain, GPT-4, FastAPI, Qdrant, Superset* **Feb 2025**

- Built internal RAG-based semantic search engine at Cyfuture using LangChain and GPT-4, enabling contextual answers from enterprise knowledge bases with natural language processing capabilities.
- Developed modular FastAPI backend for embedding generation, document chunking, and retrieval with vector search via Qdrant database.
- Integrated Apache Superset for comprehensive analytics tracking including query latency, usage patterns, and token consumption across teams.

Distributed Media Upload System | *Python, FastAPI, Kubernetes, gRPC, GridFS* **Apr 2025**

- Designed distributed microservices architecture for large file uploads using **FastAPI**, **gRPC**, **RabbitMQ**, and **GridFS** (up to 2GB files, 1,000+ users).
- Built **API Gateway** with **JWT auth**, deployed via **Docker** and **K8s StatefulSets** with auto-scaling.
- **GitHub:** `Distributed-Media-Service`

Education

Maharshi Dayanand University **November 2020 – June 2023**

Bachelor of Vocation in Software Development — GPA: 8.6/10 *Haryana, India*

Achievements & Certifications

- **Enterprise Impact:** Successfully delivered ERP system to **4,000+ government users**, approved by Ministry of Jal Shakti and validated through central-level audits.
- **Technical Excellence:**
  - **LeetCode** – Solved **440+** problems solved demonstrating strong Data Structures and Algorithms expertise
  - **GeeksforGeeks** – Achieved **Rank 2**
  - **CodeStudio** – **Level 5 Champion**
- **Certifications**
  - **Generative AI: Evolution of Thoughtful Online Search** (LinkedIn, Jul 2025)
  - **What Is Generative AI?** (LinkedIn, Jul 2025)