Yashraj Sawant

Software Developer

yashrajsawant0110@gmail.com | +91 9892344642 | Bangalore,IN | yashrajsawant.com

PROFILE

Backend & AI Application Engineer (**1.5+ YOE**) building **production LLM systems** with **Python, FastAPI, and LangChain**. Experienced in **RAG architectures**, **async inference pipelines** (Celery/Redis), and **vector search** (pgvector). Shipped AI features to **2,000+ users** while reducing infrastructure costs by **25%**.

PROFESSIONAL EXPERIENCE

Software Developer, *DSJ Keep Learning* 04/2024 – 08/2025 | Bangalore,IN

- Built automated **video intelligence pipeline** using **OpenAI Whisper API**, processing **50+ hours weekly** with chunking algorithms that **reduced transcription failures by 90%** and maintained **<200ms API latency** through async queue architecture (**FastAPI, Celery, Redis**).
- Developed **RAG system** using **LangChain** and **PostgreSQL (pgvector)** for semantic search and natural language Q&A across **5,000+ student records**, with Redis semantic caching that **reduced redundant LLM API calls by 40%**.
- Engineered **Python data pipelines** with **parallel processing** for batch transformations, **reducing execution time from 2 hours to 25 minutes** across **5,000+ records**.
- Optimized AI service costs through intelligent caching and resource management, maintaining **<2s response times** while **reducing overall operational expenses by 25%**.
- Deployed and scaled AI services on **AWS (EC2, RDS, S3)** using **Docker and Jenkins** with **auto-scaling policies**, ensuring reliable performance during high-load transcription and inference jobs.

TECHNICAL SKILLS

AI & LLM: OpenAI API, Whisper, RAG Architecture, LangChain, Prompt Engineering, Vector Search (pgvector), Semantic Caching

Backend & Cloud: Python, FastAPI, Flask, PostgreSQL, Redis, Celery, AWS (EC2, RDS, S3), Docker, Jenkins, CI/CD, React.js

PERSONAL PROJECTS

Document Intelligence System, *Tech Stack: FastAPI, LangChain, OpenAI API, PostgreSQL (pgvector), Redis*

- Built **RAG-powered Q&A system** using **LangChain** with document chunking (500-token windows) and embedding generation for **semantic search across 100+ PDFs**.
- Implemented **vector similarity search with pgvector**, achieving **sub-2s query response time** with conversation history management and context window optimization.
- Designed **Redis-based caching** for embeddings and frequent queries, **reducing OpenAI API costs by 70%** while maintaining real-time performance.

LLM-Powered Content Analyzer, *Tech Stack: FastAPI, LangChain, OpenAI API, PostgreSQL, Celery, React*

- Developed **content classification system** using **LangChain and GPT-4** to automatically categorize and extract insights from user-generated text **at scale**.
- Implemented **async processing pipeline with Celery and Redis** to handle batch LLM operations, **processing 1,000+ documents** while managing rate limits and retry logic.
- Built **analytics dashboard** showing extracted entities, sentiment analysis, and topic clusters using **React** and visualization libraries.

EDUCATION

SPPU University, *Bachelor of Engineering with a Minor in Artificial Intelligence and Machine Learning* 08/2019 – 07/2023 | Pune,IN

CERTIFICATES

- AWS Certified Solutions Architect – Associate
- IBM - Introduction to Containers w/ Docker, Kubernetes & OpenShift
- Google Cloud - Generative AI Fundamentals

LEADERSHIP EXPERIENCE

Google Developer Student Club, *Core Team Member* 08/2022 – 07/2023 | Pune,IN

- Collaborated within a 5-member core team to organize technical workshops, seminars, and hackathons, while leading a 4-member subteam to manage logistics, promotions, and outreach, driving active engagement of 100+ students in the developer community.