Shivraj Dhaytadak

+91 7972476081 | Email : **shivraj.25d@gmail.com** | Linkedin: Shivraj Dhaytadak | Github: Shivraj-Dhaytadak

Profile Summary

Dynamic **Data Scientist with 3.5 years of experience in designing and deploying Generative AI solutions**. Proven track record in optimizing workflows, reducing costs, and improving efficiency using AI/ML, Python, ETL, and cloud-native services. Skilled in **Context Engineering, RAG, and Agentic AI** to automate complex workflows and generate high-quality content.

Skills

- **Python, SQL** (Postgres , MySQL) , NoSQL (MongoDB)
- **FastAPI , LangChain, LangGraph , Crew AI ,A2A , MCP** , Pytest,Ollama, Pytorch , Transformers , vLLM
- **Agentic AI , LLM Inference , ETL , RAG, FastAPI** , REST API development , Data analysis
- **Pandas, Numpy, Scikit-learn , Pytorch ,Transformers , Seaborn, Matplotlib** , Docling , trl , OpenAI, Google genai , Hugging Face, Pydantic, Requests, BeautifulSoup, SQLAlchemy, NLTK
- **Amazon Web Services (AWS)(Amazon EC2, Amazon S3, AWS Bedrock, AWS SageMaker, Amazon RDS ,AWS lambda , AWS ECR , AWS ECS** , Amazon Fargate , Amazon CloudWatch)
- **Google Cloud Platform (GCP)** (VM Instances , Cloud Storage (Buckets), ,**Vertex AI** , Auto ML)
- **Git, Github Actions , Jenkins , Docker , Jira**

Experience

Data Scientist (Gen AI) May 2025 - Present

Yash Technologies Pune, IN

- Developed a sophisticated **Multi-agent workflow** where specialized agents collaborated to **autonomously generate, review, and refine code, improving overall code quality and consistency.**
- Engineered a **high-performance FastAPI endpoint** integrating a Text-to-SQL graph, enabling seamless natural language query translation to structured SQL **accelerating data retrieval by 40%.**
- Architected a **custom hybrid memory** framework (short-term & long-term) for chat sessions, enhancing contextual retention and summarization **boosting query accuracy by 35%** and **improved SQL success rate scores by 25%.**
- Achieved **80% reduction in manual SQL effort and 35% higher response precision**, streamlining analytics workflows and decision-making.
- Built a Multi-agent workflow using **Crew AI & Google Gemini 2.5 models** to support code generation and review across multiple programming languages.

Senior Software Engineer (Gen AI) August 2022 - May 2025

Persistent Systems Pune, IN

- Implemented high-performance model-serving endpoints using **FastAPI**, reducing **response latency by 50%** & reliable delivery for AI/ML applications.
- Created and executed automation scripts to replace repetitive tasks, enhancing precision in **data retrieval by 35%** and **boosting productivity by 25%** for product development initiatives.
- Developed a **RAG chatbot** to interact with the codebase, enabling efficient code documentation and retrieval.
- Enhanced LLM inference with RAG by leveraging document chunking, embedding optimization, and similarity search, **improving response relevancy by 40%.**
- Optimized RAG workflow by fine-tuning embeddings and retrieval strategies, **reducing latency by 30% and enhancing answer relevance.**
- Architected a data curation framework to aggregate and process datasets from multiple sources, **enabling 40% faster data integration to fine-tune LLMs, and improving the processing accuracy by 30%.**

Projects

Agentic Natural Language Query Python, FastAPI , LangGraph , Azure SQL , Azure OpenAI

- Engineered a high-performance FastAPI endpoint to invoke a Text-to-SQL graph, enabling seamless natural language query translation to structured SQL.
- Designed and implemented custom hybrid memory (short-term & long-term) to optimize chat history retention and contextual summarization, improving query accuracy and user experience.

Multi Language Code Generation Python, FastAPI , CrewAI , Gemini , Vertex AI

- Built a multi-agent workflow using CrewAI to support code generation and review across multiple programming languages.
- Enhanced agent response reliability by implementing guardrails for safety, reducing hallucinations,improving code quality consistency.

Agentic AWS Cost Estimation Python, FastAPI , LangGraph , Langchain , Gemini

- Developed an intelligent agentic system to parse cloud architecture diagrams and auto-generate AWS pricing estimates.
- Designed a dynamic questionnaire agent for configuration recommendations, enhancing resource efficiency and cost optimization.

Solar Scan - Code retrieval RAG Python, FastAPI , Gemini , Chroma , Sqlalchemy ,Pydantic ,Postgres

- Built an AI-powered chatbot to interact with large codebases, enabling efficient code search, documentation generation, and retrieval.
- Integrated Google Gemini AI with Ollama for hybrid retrieval (local + API-based inference), ensuring faster responses and improved developer productivity.

Education

Bachelor's of Engineering in Computer Engineering Pune,In
PES Modern College of Engineering , Pune August 2018 - July 2022

Certification

Azure Certified AI Fundamentals - Issued August 2025
AWS Cloud Practitioner - Amazon Web Services - Issued April 2024 - Expires April 2027

Awards

Valuable Individual Asset Award : is a direct reflection of the work we have put into developing Agentic AI workflows.
High Five Individual Award: For building multiple frameworks across teams and improving results through CI/CD.
Bravo Team Award: Guided and mentored team members, fostering a collaborative and productive environment.