

Devang Borkar

+1 (312) 358-5722 | devangborkar3@gmail.com | LinkedIn | GitHub | Portfolio

Education

- M.S in Computer Science from University of California Davis Status-Ongoing
- B.E. in Computer Science from Pune University Status - Graduated

Work History

SWE Intern at *LearnHaus AI*

June 2025 - August 2025

- Developed an end-to-end multimodal data processing pipeline using Python with async processing that orchestrated video/audio analysis speech-to-text transcription, and multi-LLM integration to deliver automated coaching.
- Built an automated feedback evaluation system using Model Context Protocol(MCP) for agentic behavior with AI-as-a-judge, improving feedback actionability.

Founding Engineer at *HammerTrade (Stealth Startup)*

October 2024 to June 2025

- Developed a high-throughput, distributed data processing service in Python to manage ML workloads for high-frequency trading simulations, ensuring performance & scalability.
- Deployed the platform on AWS using Kafka for robust data ingestion and processing of real-time market data streams.
- Engineered a complex market simulation environment to train autonomous reinforcement learning (RL) agents, modeling extreme volatility with over 10 configurable parameters.

Software Engineer at *Hexaview Technologies*

August 2022 to September 2024

- Shipped 20+ features for a Fortune 500 wealth management platform, developing a scalable backend using **ASP.NET Core** with **AWS Lambda-based microservice** architecture.
- Maintained a legacy backend servicing over 1 million monthly requests, applying key design patterns to reduce code complexity & successfully redesigning 50+ REST APIs.

Projects

ResChat – Decentralized Platform with AI Assistant

- Built a low latency communication platform using C++ and Python leveraging distributed storage systems for real-time messaging and large file transfers
- Implemented a RAG-based AI chatbot using LangChain for document parsing across distributed databases and reducing information retrieval time by 85%.
- Developed a pipeline to generate high-quality embeddings and index documents in a FAISS vector database, optimizing for accurate embedding-based retrieval.

Gitartha Engine – Semantic Search for the Bhagavad Gita

- Architected a full-stack application using Go (Gin) for the high-concurrency REST API and FastAPI for ML model inference, achieving consistent P95 search latency of under 15ms.

- Developed low latency semantic search using PostgreSQL with the pgvector extension, resulting in an average query response time of 12.7ms across a corpus of 700+ verses.

LLM Self-Chat - Agentic AI Simulation Framework

- Developed an agentic framework using Python and LangChain enabling multiple LLMs to converse, simulating multi-agent systems for behavior analysis and prompt engineering.
- Engineered the system with custom prompts for each agent and utilized WebSockets for real-time, low-latency communication between the front-end and a Flask backend.

Daily Digest – AI-Powered Gmail/Calendar Summarizer

- Created an AI assistant reducing the daily planning overhead by 70% using Flask and Python powered by Gemini AI via secure OAuth 2.0, providing personalized priority-based summaries and Text-To-Speech capabilities.

AI CodeMentor – LLM-Powered Code Analysis & Review Automation

- Created a GitHub Action using Node.js for automated CI/CD code reviews, supporting PRs, issues with intelligent change detection via git diff, GitHub API, and fallback mechanisms.
- Added agentic tool calling features with OpenAI APIs to enable the LLM to dynamically invoke external functions and tools for extended analysis during code reviews.