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

AI Engineer at *PilotCrew AI* — October 2025 - Present

- Built a Python-based LLM evaluation orchestrator (FastAPI, MongoDB, Redis) that discovers actionable "failure modes" and generates targeted adversarial edge-case tests.
- Developed a closed-loop alignment workflow that converts failures into corrective prompt updates improving agent performance by up to 54% on adversarial eval runs.

SWE Intern at *LearnHaus AI* — June 2025 - August 2025

- Engineered a 0-to-1 multimodal coaching service in React & Python that orchestrates asynchronous video, audio, and text processing enabling users to improve public speaking capabilities through an automated, latency-optimized feedback system
- Utilized LLM-As-Judge with multi-provider integration to automate ground-truth generation ensuring evaluation reliability without manual labeling and deployed the platform to GCP

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.
- 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.
- Optimized a legacy C# backend servicing over 1 million monthly requests, applying key design patterns to reduce code complexity & successfully redesigning 50+ REST APIs.

Projects

CausalFlow – Autonomous Agent Debugging Framework

- Built an interpretable agentic framework achieving 40% performance uplift over baseline to resolve failures in multi-step reasoning chains for long horizon complex tasks
- Engineered deterministic synthetic environments to ground agent execution in verifiable state transitions, eliminating hallucination risks associated with LLM-based world modeling

Process Reward Model (PRM) for On-Device LLMs

- Built an composite inference system coupling a lightweight generator (Qwen3-0.6B) with a heavy verifier (Qwen3-8B), enabling efficient “weak-to-strong” generalization for resource-constrained environments
- Achieved a 21% performance uplift over self-consistency baselines on the GSM8K benchmark by engineering a PRM-guided Best-of-N search strategy, effectively mitigating logical hallucinations in sub-1B parameter models.

AI CodeMentor – LLM-Powered Code Analysis & Review Automation

- Developed an LLM-powered agent for automated CI/CD code reviews, using agentic tool calling (OpenAI APIs) and Node.js to analyze PRs and issues.
- Engineered the agent to parse git diffs via the GitHub API and invoke external analysis functions, providing intelligent, context-aware feedback on code changes.

LLM Self-Chat - Agentic AI Simulation Framework

- Built an agentic framework using Python and LangChain, enabling multi-agent LLM simulations for behavior analysis and prompt engineering.
- Integrated WebSockets to establish a real-time, low-latency communication channel between the React front-end and Flask backend for interactive agent simulation.

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.

Daily Digest – AI-Powered Gmail/Calendar Summarizer

- Built 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.