Pavan Sai Reddy Pendry

pavansaipendry2002@gmail.com | (785) 813-7825 | Irving, TX
pavansaipendry.dev | github.com/pavansaipendry | linkedin.com/in/pavansaireddypendry

PROFILE

Software Engineer specializing in Generative AI, LLMs, and RAG systems with production experience shipping multi-agent applications and async backend services. Built BabyJay, a production AI campus assistant with multi-stage RAG serving 7,300+ courses (82.4% user approval), and AI City, an autonomous multi-agent simulation orchestrating 10+ LLM agents. Strong across Python, FastAPI, Claude/GPT, PyTorch, Docker, PostgreSQL, and vector DBs. Published researcher (Springer, IEEE), M.S. CS May 2026.

TECHNICAL SKILLS

Languages: Python, Java, TypeScript, JavaScript, C++, SQL
AI/ML & GenAI: LLMs (Claude, GPT-4, Gemini, Llama), Agentic AI, Multi-Agent Systems, RAG, Prompt Engineering, LangChain, LangGraph, Hugging Face Transformers, PyTorch, scikit-learn, NumPy, Pandas
Vector DBs: ChromaDB, Qdrant, pgvector
Backend & APIs: FastAPI, Node.js, Express, REST, WebSocket, JWT, async Python
Data & Infra: PostgreSQL, Redis, Supabase, Docker, AWS, Linux
Practices & Tools: Git, CI/CD (GitHub Actions), Pytest, Agile, system design, unit and integration testing, monitoring

EXPERIENCE

University of Kansas | Lawrence, KS **Jan 2025 – May 2026**
Research Software Engineer

- Built **BabyJay** (babyjay.bot), a production AI campus assistant serving 7,300+ courses, 2,207 faculty, dining, transit, and EECS programs, achieving **82.4% user approval** on real student feedback collected through a thumbs up and thumbs down pipeline.
- Designed a **FastAPI** backend with 14+ authenticated endpoints, **Supabase PostgreSQL** persistence for conversations, messages, and feedback, **JWT** auth with per-user isolation, and a 3-tier rate limiter capped at a daily cost budget.
- Engineered a multi-stage **RAG** pipeline with a preprocessor, classifier, router, **8 specialized retrievers**, and a context builder feeding Claude, reducing average retrieval from 500 to 1000ms down to 5 to 50ms, a **35x improvement** over pure vector search.
- Built **9** production Python web scrapers collecting course catalog, faculty directory, dining, and GTFS transit data with retry logic, deduplication, and schema validation, plus automated **Pytest** suites for retrieval accuracy and CI regression checks.
- Deployed the backend on **Render** and the React frontend on **Vercel** at the custom domain **babyjay.bot**.

Note | USA, Remote **May 2025 – Aug 2025**
Software Engineer Intern

- Built **Note**, a developer intelligence platform capturing and analyzing Claude Code sessions, designing and implementing **25+ REST API** endpoints in **Next.js 16** App Router covering auth, prompts, projects, search, analytics, and cross-session intelligence features.
- Designed a normalized **PostgreSQL** schema with 15 tables (prompts, sessions, projects, knowledge graph, audit log) using **tsvector** with **pg_trgm** trigram similarity for fuzzy search, composite indexes, and auto-updating search vectors via triggers.
- Built a **WebSocket** server with **Redis** pub/sub for real-time CLI-to-web session pairing using hashed 6-digit codes, plus **JWT** dual-token auth (7d access, 30d refresh) with bcrypt, token revocation, and rate limiting via express-rate-limit.
- Built a Node.js CLI with 24 commands (save, search, standup, report, capture, knowledge, share) and a React dashboard for session analytics and search.

Amrita Vishwa Vidyapeetham | Kerala, India **Jun 2023 – May 2024**
Research Assistant

- Co-authored **2 peer-reviewed papers** in **Springer LNNS** ICT4SD 2024, DishKit and **IEEE i-PACT 2023** Sweep Spot, and shipped **2 production apps** in **Python**, **React**, and **PostgreSQL** serving 500+ users with REST APIs, JWT auth, and **AWS** in a 6-person Agile team.

PROJECTS

FinDocAgent | *Python, PyTorch, Hugging Face, LangChain, LangGraph, FastAPI, Docker*

- Fine-tuned **DistilBERT** on SEC filings (10-K, 10-Q) using **PyTorch** and **Hugging Face Transformers** to classify document sections, hitting **92%** accuracy for downstream routing.
- Built a multi-agent pipeline with **LangGraph** (Parser, Retriever, Analyzer) that produces cited answers from filings, with **LangChain** handling **RAG** over a **pgvector** store.

AttentionFM | *SvelteKit, FastAPI, async Python, WebSocket, Claude Sonnet 4, Docker, RunPod*

- Built a 24/7 AI podcast platform with a fully async **FastAPI** backend handling concurrent room management, **WebSocket** multiplexing across connected clients, and a bidirectional protocol handling 7 event types with base64 audio payloads.
- Containerized the platform with **Docker Compose** across backend, PostgreSQL, Redis, and Qdrant, built a **RunPod** serverless handler for GPU workloads, and shipped a **SvelteKit** frontend with real-time transcript streaming and Web Audio API playback.

AI City | *Python, FastAPI, TypeScript, PixiJS v8, PostgreSQL, Redis, Qdrant, LangGraph*

- Built an autonomous multi-agent system using **LangGraph** orchestrating **10+** LLM agents with stateful workflows, conditional routing, and persistent memory via **Qdrant**, demonstrating production patterns for Agentic AI frameworks.
- Designed a 10-migration **PostgreSQL** schema, **Redis** messaging, and Qdrant vector memory backing each agent, with an isometric **PixiJS** frontend visualizing agent decisions and interactions in real time.

HACKATHONS

HackKU 2025 | University of Kansas ACM, 36-Hour Hackathon **Apr 2025**

- Built and shipped a cloud-based AI-powered application in 36 hours using **Claude API**, **Python**, and **TypeScript** with clean code, automated tests, and Agile practices, earning recognition for the most innovative use of AI among 60+ competing teams.

Hack K-State 2025 | Kansas State University, 36-Hour Hackathon **Oct 2025**

- Shipped a real-time collaborative dashboard with **Python** back-end, **WebSocket** streaming, and **React** front-end that automated cross-functional team task routing using AI classification.

EDUCATION

University of Kansas | Lawrence, KS **Aug 2024 – May 2026**
M.S. Computer Science
Coursework: **Software Engineering**, **Algorithms**, **Database Systems**, **Distributed Systems**, **Computer Architecture**, **Machine Learning**

Amrita Vishwa Vidyapeetham | India **Oct 2020 – May 2024**
B.Tech. Computer Science and Engineering
Coursework: **Data Structures**, **Algorithms**, **Operating Systems**, **Computer Networks**, OOP (Java, C++), Cloud Computing, Deep Learning, Web Development