Jay Kilaparthi

New York, NY | +1 (212) 729-5295 | jayakeerthk@gmail.com | www.linkedin.com/in/jayvk | jayvk.com

SUMMARY

Software Engineer & Product Leader with 3+ years of experience leading GenAI systems across startups and applied research. Engineered LLM-powered agents for voice, email, and search, combining real-time pipelines with scalable, serverless infrastructure.

- Specialized in RAG pipelines, Model Context Protocol (MCP), and agentic AI workflows.
- Hands-on with OpenAI, Hugging Face, LangChain, and vector-based retrieval using FAISS and Neo4j.
- Deployed scalable ML systems using AWS (SageMaker, S3, Lambda, EC2), Docker, and MLOps pipelines.
- Built secure API integrations with OAuth, Twilio, and Gmail, optimizing for speed, privacy, and reliability.
- Experience with data analytics, strategy consulting, and nonprofit operations before transitioning into AI leadership and startups.

SKILLS

GenAI & NLP: RAG Pipelines, Model Context Protocol (MCP), Agentic AI, Prompt Engineering, LangChain Agents, LLM APIs (OpenAI, Gemini), Whisper, Vector Search, Fine-tuning, Knowledge Graphs
AI Engineering: LangChain, Hugging Face, TensorFlow, Keras, PyTorch, Embedding Models, Neo4j (Graph RAG), ML/LLMOps, Lightweight LLMs, LoRA/QLoRA, Transformers
Infrastructure & Ops: AWS (SageMaker, S3, Lambda, EC2), Docker, Firebase, OAuth, CI/CD, Airflow, Snowflake, Databricks
Data & Product: SQL, NoSQL, Data Pipelines, n8n, Product Strategy, Jira, Confluence, Cross-Functional Collaboration
Frontend/Backend: Python, JavaScript (Node.js, React), HTML/CSS, REST APIs, Secure API Integrations, System Design
Other: Excel, Tableau, Power BI, Alteryx, Git & GitHub, Agile & Scrum

EDUCATION

Master of Science – Information Systems
Baruch College Zicklin School of Business | New York City

- CUNY Reading Corps tutor for NYC public schools; supported early-grade students through structured fluency sessions and tracked weekly literacy progress as part of a citywide academic intervention.
- Led the development of SafeBite NYC, a public health analytics project using Python and NYC inspection data to surface high-risk restaurant zones through visual filtering and clustering.
- Conducted IT strategy assessment for Baruch College's LMS migration (Blackboard → Brightspace), delivering recommendations on change management, system usability, and stakeholder communication.

PROJECTS & RESEARCH

Fine-Tuning Lightweight LLMs for Code Generation
Python Library (aicoderx 1.1.0) | Released on Hugging Face & PyPI

- Fine-tuned a 0.5B open-source LLM (Qwen2.5-Coder) using QLoRA and PEFT adapters on AWS SageMaker with Hugging Face containers, targeting domain-specific Python code generation.
- Quantized the model to Q5_K_S and achieved sub–1s inference latency on CPU-only devices, reducing memory usage by 3x and enabling local deployment without GPU dependency.
- Packaged the training pipeline and inference engine into aicoderx, a pip-installable Python library for offline code generation and edge AI workflows.
- Integrated CI/CD workflows with GitHub Actions and versioned model checkpoints via S3, enabling reproducible training, inference testing, and scalable release cycles.

TensorFlow & Computer Vision
AI-Based Fault Detection in Solar Panels | Patent Published (India)

- Built a deep learning pipeline using TensorFlow and Keras to classify solar panel faults from I–V and P–V curve images, achieving 94%+ accuracy on real-world degraded signal data.
- Engineered end-to-end ML system: custom data preprocessing, domain-specific augmentations, model training, evaluation, and deployment-ready inference module.
- Developed modular architecture enabling scalable integration into solar diagnostics workflows; validated on multiple panel types and field scenarios.
- Co-inventor on published patent covering the full AI-based system for solar fault detection, awarded by the Indian Intellectual Property Office in 2023.

EXPERIENCE

Founding Engineer GenAI

Keeya Labs | New York City | Jan 2025 – Present

Current: Voice-first email assistant using LLMs to summarize Gmail and manage inbox via SMS, voice, and chat

- Designing and deploying a fully voice-native GenAI system using GPT-4o, LangChain, and AWS (Lambda, S3) to enable inbox summarization, search, and control over Twilio Voice/SMS
- Building a RAG pipeline with hybrid keyword–semantic search, contextual compression, and memory modules to support multi-turn conversational flows across long email threads
- Engineering OAuth2-based Gmail integration with token-scoped access and secure backend isolation to ensure compliant access across voice and SMS endpoints
- Implementing server-side caching for vectorized email embeddings and prompt templates to optimize latency and enable reusable context across sessions
- Tuning LLM prompt templates and context handling to improve summarization fidelity, reduce hallucinations, and support noisy, real-world inbox data
- Leading product roadmap, prompt strategy, and infrastructure planning to enable voice-based LLM actions like reading, summarizing, and searching inbox contents

Launched: Voice-powered storytelling platform preserving memories via GPT and voice cloning APIs

- Developed real-time voice cloning pipeline using Whisper and ElevenLabs APIs, achieving 90%+ voice preservation accuracy on memory clips up to 10 minutes.
- Built a RAG-based story generation system with OpenAI/GPT and Hugging Face models, allowing users to interact with AI personas of loved ones.
- Reduced inference latency by 40% via model quantization, prompt optimization, and deployment on cost-efficient GPU instances.
- Designed and deployed full-stack system using Next.js, AWS (Lambda, S3), and Supabase for audio storage, prompt orchestration, and session control.
- Supported 500+ memory submissions in MVP stage; implemented Stripe-powered credit system and iterated pricing model through A/B tested user flows.

Co-Founder & Technical Lead

Patchly | New York City | October 2024 – March 2025

Campus startup focused on improving student event discovery and engagement across CUNY colleges

- Built and launched the MVP using Firebase, React, and metadata filtering to recommend events by location, interest, and time.
- Built backend workflows for real-time event submission, approval, and recommendation by interest and location.
- Managed technical roadmap, testing cycles, and team operations across multiple CUNY campuses.
- Won the CUNY New Venture Accelerator Case Competition and secured $1,250 in early-stage funding for MVP.

Graduate Teaching Assistant

Baruch College | New York City | September 2024 – May 2025

Assisted faculty in delivering a graduate-level Python analytics course

- Supported 150+ students on Python, NumPy, Pandas, and data visualization for real-world business cases.
- Provided 1:1 guidance and graded projects, reinforcing applied data analysis skills and debugging workflows.
- Collaborated with faculty to iterate course materials and improve learning outcomes based on student feedback.

Early Leadership & Strategy Experience

Multiple | India | August 2021 – December 2023

Worked across nonprofit fundraising and consulting roles, leading education-focused campaigns

- Managed a 15-person volunteer team, exceeding fundraising targets by 20% and expanding education access to 100+ underserved students across India.
- Built Power BI dashboards and streamlined donor tracking workflows across regional teams in multiple cities.
- Led volunteer onboarding and drove cross-functional alignment across education, outreach, and fundraising teams, accelerating campaign timelines and execution.
- Supported strategy consulting at Impact Consulting by conducting stakeholder interviews, synthesizing insights, and building nonprofit growth frameworks.
- Delivered executive-facing slide decks and analytics reports that shaped fundraising strategies.