JAY PATEL

Chicago, IL - 60016 • +1(224) 266-3259 • jaykrinapatel@gmail.com
• Github: https://github.com/jaypatel0112 • Linkedin: https://www.linkedin.com/in/jay-patel-605a711a0 •

SUMMARY

AI Software Engineer with 2+ years building LLM-powered applications, agentic systems, and cloud-native backend services. Experienced in designing multi-agent orchestration, RAG pipelines, and ReAct-style agentic patterns with full observability across production environments. Skilled in prompt engineering, REST API development, and deploying high-throughput distributed architectures on AWS using Python.

SKILLS

- **Languages:** Python, Java, C#, JavaScript, SQL
- **AI/ML:** LLMs, Prompt Engineering, RAG, Fine-Tuning, Multi-Agent Systems, LangGraph, LangChain, ReAct Patterns, Tool-Augmented LLMs, HuggingFace Transformers, Rasa NLP/NLU
- **Orchestration:** Agentic Pipelines, MCP Servers, Tool Use / Function Calling, State Management
- **Observability:** LLM Tracing, Token Optimization, Cost-per-Request Metrics, Distributed Logging, Performance Benchmarking
- **Cloud:** AWS (EC2, S3, Lambda, CloudWatch), GCP (Vertex AI), Docker, Jenkins, CI/CD
- **Systems:** Distributed Systems, Microservices, REST APIs, ASP.NET, Node.js, High-Throughput Architectures
- **Data:** Elasticsearch, OpenSearch, PostgreSQL, MySQL, MongoDB, Large-Scale Processing

EXPERIENCE

Full Stack AI Developer | Peterson Technology Partners, Chicago, Illinois — April 2025 – Current

- Architected Agentic AI systems to autonomously execute recruiting workflows sourcing, screening, ranking, and shortlisting candidates reducing manual recruiter workload with minimal human intervention.
- Designed production-grade LLM pipelines with tool use, state management, and structured output parsing across high-concurrency production environments.
- Built agentic systems using ReAct-style patterns enabling LLMs to reason, call external APIs, and iterate toward task completion with full observability hooks for tracing and cost monitoring.
- Developed RAG pipelines over Elasticsearch/OpenSearch (1M+ records), improving retrieval accuracy and query latency for LLM-grounded responses at scale.
- Deployed scalable REST APIs on AWS (EC2, Lambda, CloudWatch) supporting high-concurrency LLM workloads with fault tolerance and real-time performance monitoring.

Software Developer Intern | Advantage IT Inc, Hillsboro, OR — August, 2024 – April, 2025

- Collaborated with senior developers to design and implement software solutions, contributing to coding, testing, debugging, and third-party API integrations across the full development lifecycle.
- Researched and integrated AI agent tools to enhance application automation capabilities, and contributed to early-stage Agentic AI workflows within existing development pipelines.
- Participated in code reviews promoting clean code practices, and supported Agile processes through sprint planning, stand-ups, and technical documentation.

Software Engineer Intern | Kintu Designs Pvt Ltd, Surat, India — Sep 2021 – May 2022

- Developed a Rasa NLP/NLU-based healthcare chatbot automating doctor appointment scheduling, handling 500+ daily interactions with 95% success rate and reducing patient wait time by 50%.
- Designed and optimized REST APIs for concurrent user load, improving response time and system reliability across multiple services.
- Improved CI/CD pipelines for 30+ microservices using Jenkins and Docker, increasing deployment reliability by 25%.

EDUCATION

DePaul University – Chicago, IL
Master of Science in Computer Science — GPA - 3.78/4

Charotar University of Science and Technology - Anand, India
Bachelor of Technology in Computer Science and Engineering — GPA - 9.35/10

PROJECTS

Course Enrollment & Event Portal | *Springboot, SQL, Azure, MongoDB* — Source code

- Built a distributed REST API system supporting 50K+ students, optimized for high read/write and concurrent usage.

Grocery Mobile Application | Lead Developer (Team of 3) | *Android Studio, Java* — Source code

- Developed a data-driven Android application with real-time pricing, recommendation logic, and scalable APIs.