SAI TEJA MEKA

(860) 494-9460 | saitejameka45usa@gmail.com | Portfolio | LinkedIn | New Jersey, USA

PROFESSIONAL SUMMARY

Production AI builder with 1+ years of experience in engineering LLM systems, multi-agent orchestration, and full-stack AI applications. Deployed RAG pipelines achieving 85% consistency with human evaluation. Specialize in translating research into shipped products—from FastAPI backends to real-time inference systems, seeking frontier AI engineering roles building transformative applications.

SKILLS

- **AI/ML Engineering:** Multi-agent systems • LLM orchestration • RAG pipelines • Prompt engineering • Vector databases (ChromaDB, Pinecone) • Model evaluation frameworks • Fine-tuning • Embeddings • Semantic search
- **Backend & APIs:** Python • FastAPI • Node.js • REST APIs • WebSockets • JWT authentication • Rate limiting • Real-time streaming • Async processing
- **Frontend & Visualization:** React • Next.js • TypeScript • React Three Fiber • WebGL • Framer Motion • Streamlit • Real-time UI
- **Data & Infrastructure:** PyTorch • TensorFlow • LangChain • Crew AI • OpenAI/Anthropic APIs • Docker • Git • SQL • NumPy • Scikit-learn
- **Cloud & DevOps:** AWS (EC2, S3, Lambda) • GCP • Vertex AI • CI/CD • Container orchestration • Production deployment

PROJECT EXPERIENCE

Deep Blue: Production Multi-Agent AI Platform

FastAPI • Next.js • OpenAI API • ChromaDB • RAG • JWT Auth • Real-time Streaming

- Architected production-grade AI platform with sophisticated multi-agent orchestration coordinating 5+ specialized agents (research, code analysis, document processing, email, calendar) through intelligent task routing and state management—demonstrating 0-to-1 system design capability
- Engineered RAG-powered memory system using ChromaDB vector database with semantic search, achieving context-aware responses across unlimited conversation threads with sub-200ms retrieval latency and per-session memory isolation
- Built full-stack application: FastAPI backend (JWT authentication, rate limiting, error handling, async processing) + Next.js/React frontend with real-time streaming responses via WebSockets, voice input (Web Speech API), multi-format file processing (PDF/DOCX/CSV)
- Implemented production-ready features: conversation tagging/search, multi-format exports (Markdown/JSON/TXT), intelligent filtering, Framer Motion animations matching Claude/ChatGPT/Perplexity UI standards—deployed with Docker containerization

LLM Content Classification & Evaluation System | GitHub ↗ | Live Demo ↗

GPT-4 • Claude • LangChain • Python • Hugging Face • Scikit-learn • Prompt Engineering

- Built a modular agent-driven content moderation pipeline using GPT-4 and Claude APIs with few-shot and chain-of-thought prompting, achieving 85% consistency with human review across 10,000+ video transcript segments (safe-for-ads, satirical, explicit classifications)
- Engineered multi-model evaluation framework with LangChain orchestration—automated side-by-side prompt template comparison and policy compliance verification for streaming/advertising contexts
- Designed adaptive decision logic simulating real-world compliance scenarios—handling dynamic policy changes and contextual viewer standards with structured output validation and confidence scoring

ScholarGPT: Modular AI Learning System

Crew AI • OpenAI API • Web Speech API • Browser Extensions • Agent Orchestration

- Designed agent-powered study platform leveraging Crew AI's orchestration framework with specialized agents (summarizer, quiz generator, concept explainer) for rapid experimentation and personalized learning paths
- Engineered browser-native workflows with Web Speech API for voice-driven note summarization, instant flashcard generation, and adaptive concept deep-dives—enabling real-time content transformation
- Implemented modular architecture supporting user-driven agent customization and workflow extension—transforming static study tools into evolving, context-aware learning companion

AI Resource Optimization Simulator

Python • NumPy • Scikit-learn • Matplotlib • Streamlit • Predictive Modeling

- Developed probabilistic forecasting simulator modelling renewable energy generation (solar, wind, hydro, battery) with ML-enhanced resource balancing algorithms optimizing output under variable weather conditions
- Built rule-based and ML hybrid forecasting system with Streamlit dashboard visualizing energy allocation strategies and storage optimization—demonstrating practical application of predictive analytics

PROFESSIONAL EXPERIENCE

AI Engineer Intern | Cloud Bridge Solutions Inc, Marlborough, MA
Sep 2025 – Present

- Architected enterprise-scale AI education platform with plugin-based microservices architecture—engineered 9 production-ready modules (RAG-powered tutor, adaptive testing, lesson generation, analytics) across 15,000+ lines of Python/JavaScript, implementing event-driven communication with Flask Blueprint system and MongoDB document models supporting 18+ collections for multi-tenant data isolation
- Built RAG engine achieving sub-250ms semantic retrieval with Pinecone vector database—implemented document chunking pipeline processing 500+ educational resources, OpenAI embedding generation with metadata filtering (grade/subject/difficulty), and context assembly system maintaining 92%+ relevance scores in development testing across 1,200+ query scenarios
- Engineered AI content generation services with GPT-4 API integration—developed lesson planner with curriculum standards alignment generating structured outputs (objectives, activities, assessments), adaptive quiz system with difficulty progression algorithms, and automated grading logic, achieving 88% accuracy against teacher-validated rubrics in internal evaluation

Data Science Consultant | PepsiCo Capstone Graduate Project, Valhalla, NY
Aug 2024 – Dec 2024

- Developed predictive ML models optimizing packaging performance and material waste reduction aligned with sustainability objectives—building statistical simulation frameworks that accelerated R&D decision velocity by 25%
- Designed AI-driven experimental frameworks evaluating packaging configurations through automated testing pipelines—enabling faster data-driven product development cycles with reproducible methodologies

Data Analyst | Savant Instruments Pvt. Ltd, Hyderabad, India
July 2022 – July 2023

- Engineered real-time Power BI dashboards and AI-powered reporting pipelines, reducing manual reporting overhead by 40% while delivering campaign performance insights with sub-hour latency
- Automated Python/SQL ETL workflows and built forecasting engines, improving prediction accuracy by 25% with end-to-end experiment tracking and version control for reproducible analytics
- Conducted A/B testing experiments optimizing campaign strategies—bridging data insights with business outcomes through systematic hypothesis testing and statistical validation

EDUCATION

Master of Science in Business Analytics and Project Management | University of Connecticut, Hartford, CT
Aug 2023 – May 2025 | Focus: Data Science & AI Systems | **Relevant Coursework:** Generative AI, Statistics, Data Science (Python), Big Data & Cloud, Predictive Modeling, Data Mining, Business Process Modeling, Text Mining, Time Series Forecasting, Project Management

Bachelor of Engineering in Electronics and Communication | VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
Aug 2018 – July 2022

CERTIFICATIONS & RECOGNITION

AWS Machine Learning Certification | IBM AI Engineering Professional Certificate | Active Open-Source Contributor