SAI TEJA MEKA

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

PROFESSIONAL SUMMARY

AI engineer with 1+ years of experience building LLM systems, multi-agent orchestration, and full-stack AI applications. Developed RAG pipelines achieving 85% consistency with human evaluation. Built novel AI infrastructure including an LLM reasoning debugger (Chronos) and a cognitive architecture extraction engine for synthetic expert personas. Specialize in translating research into shipped products—from event-sourced backends to real-time 3D visualization systems.

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, Cognitive architecture extraction, Big Five personality modeling, LLM observability
- **Backend & APIs:** Python, FastAPI, Node.js, REST APIs, WebSockets, JWT authentication, Rate limiting, Real-time streaming, Async processing, Event sourcing, PostgreSQL, Neo4j, MongoDB, Async processing
- **Frontend & Visualization:** React, Next.js, TypeScript, React Three Fiber, WebGL, Framer Motion, Streamlit, Real-time UI, Three.js, GLSL shaders, React Flow, Tailwind CSS
- **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

Chronos: LLM Time-Travel Debugger | GitHub ↗ | Live Demo ↗
AWS EC2 • FastAPI • PostgreSQL • React Flow • Event Sourcing • OpenAI/Anthropic-ready • SSE

- Architected an event-sourced LLM observability platform that captures step-level traces (messages, tool calls/results, errors) and enables time-travel debugging: rewind to any state, fork "what-if" branches, and compare timelines to isolate divergence points.
- Built an interactive conversation DAG UI in React Flow with live updates via server-sent events (SSE), making failures inspectable and branch differences explorable through compare + timeline playback + Arena metrics.
- Deployed the full stack on AWS EC2 using Docker Compose (backend, Postgres, Redis, pgAdmin) with health checks and restart policies for reliability, exposing the API on port 8000 for end-to-end testing.
- Currently validating on AIMO3 mathematical reasoning benchmarks to demonstrate practical value in isolating multi-step LLM reasoning failures

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

Agent Persona Engine | GitHub ↗ | Live Demo ↗
FastAPI • React • ChromaDB • Neo4j • OpenAI API • WebSockets • Big Five Personality Model

- Solo-built cognitive architecture extraction engine creating synthetic expert personas from technical discussions—extracting 8 layers of cognitive DNA (identity, mental models, reflexes, reasoning chains, Big Five personality traits) architected for improved domain-specific performance over vanilla LLMs
- Engineered hybrid memory system combining ChromaDB vector search with Neo4j knowledge graph, achieving context-aware responses with sub-200ms retrieval and personality drift detection with auto-correction.
- Built production React frontend + FastAPI backend with real-time WebSocket streaming, knowledge graph visualization, and side-by-side persona vs. vanilla GPT comparison demonstrating measurable domain expertise improvement

3D Interactive Portfolio Website | GitHub ↗ | Live Demo ↗
React Three Fiber • Three.js • TypeScript • GLSL Shaders • Zustand • Groq API • Tailwind CSS

- Engineered immersive 3D portfolio using React Three Fiber with custom GLSL shaders, GPU particle systems, and cinematic postprocessing (bloom, chromatic aberration, vignette)—delivering 60fps WebGL experience.
- Integrated an AI assistant (Groq API) with a terminal-style UI to answer questions about projects/skills, reducing "time-to-info" for recruiters by enabling in-page Q&A.
- Implemented interaction + navigation plumbing (raycast selection, smooth camera transitions, deep linking) to make 3D projects discoverable and usable like a standard website.

CURRENT WORK

- ***Shipping:*** Pioneering a novel AI reasoning architecture that fuses LLM observability infrastructure with synthetic expert consultation — enabling AI systems to autonomously recruit domain-specialist personas mid-reasoning, achieving vertical depth scaling without compute-intensive horizontal expansion
- ***Researching:*** Developing psychometrically grounded persona generation framework leveraging Big Five personality trait modeling to produce synthetic experts with empirically validated behavioral profiles — bridging computational psychology with applied AI systems
- ***Exploring:*** Designing recursive metacognitive architectures where AI systems leverage their own observability telemetry to autonomously identify reasoning failures, consult domain experts, and self-optimize across conversation timelines — a step toward self-improving inference systems

PROFESSIONAL EXPERIENCE

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

- Architected enterprise-scale AI education platform with plugin-based microservices—engineered 9 production-ready modules (RAG-powered tutor, adaptive testing, lesson generation, analytics) across 15,000+ lines of Python/JavaScript using Flask Blueprint and MongoDB with 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 with OpenAI embedding generation and metadata filtering (grade/subject/difficulty), maintaining 92%+ relevance scores across 1,200+ query scenarios
- Engineered AI content generation services with GPT-4 API integration—developed lesson planner with curriculum standards alignment, adaptive quiz system with difficulty progression algorithms, and automated grading logic achieving 88% accuracy against teacher-validated rubrics
- Authored technical documentation including API endpoint references and module integration specifications across the 9-module microservices platform, supporting cross-team development and onboarding
- Collaborated cross-functionally with engineers and product stakeholders on feature development priorities using agile workflows in ClickUp, aligning deliverables with educational content standards across modules
- Participated in code reviews and conducted validation testing across the RAG retrieval pipeline and content generation services, ensuring consistent quality across curriculum updates

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

- Developed predictive ML models (linear regression, gradient-boosted trees) optimizing packaging performance—applied feature engineering on material properties and environmental variables, improving prediction accuracy by reducing error margins and enhancing R^2 values across complex multi-variable datasets
- Delivered a simulation tool enabling researchers to test packaging recipe outcomes via automated virtual experimentation rather than physical prototyping, accelerating R&D decision velocity by 25%

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

- Engineered real-time Power BI dashboards and AI-powered reporting pipelines for a B2B water analytics company serving industrial, laboratory, and government sectors across southern India—reducing manual reporting overhead by 40% with sub-hour latency insights
- 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
- Designed and executed A/B testing experiments optimizing B2B marketing campaigns across diverse client segments—serving as lead analyst overseeing experiment design, statistical hypothesis testing, and performance reporting that directly informed campaign strategy adjustments

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 Certified AI Practitioner (Amazon Web Services, issued Sep 2025, valid through Sep 2028) | IBM AI Foundations Coursework — 4-course series via Coursera (IBM, completed Aug 2020) | Active Open-Source Contributor