Krishna Chaitanya Bodepudi

St. Louis, MO (Open to Relocate) | 6468211711 | kcbodepudi@gmail.com | linkedin.com/in/krishna-chaitanya-bodepudi

APPLIED AI / ML ENGINEER (PRODUCTION ML & LLM SYSTEMS)

Applied AI / ML Engineer with hands-on experience building **production-oriented ML and LLM-backed systems** in small-to-mid scale environments. Focused on correctness-first system design, including **request scoping, data isolation, and predictable failure behavior**. Experienced in developing **FastAPI-based inference services, semantic retrieval pipelines, and persistent memory architectures** with practical exposure to deploying, operating, and maintaining these systems beyond experimental prototypes.

TECHNICAL SKILLS

Languages: Python, SQL
Machine Learning: PyTorch, Scikit-learn, Feature Engineering, Model Training & Evaluation
LLM & NLP Systems: Hugging Face Transformers, RAG Architectures, Embeddings, Vector Search
Inference & Deployment: FastAPI, REST APIs, Async Processing, Latency Optimization
MLOps: Docker, CI/CD (GitHub Actions)
Data & Storage: PostgreSQL, SQLite, MongoDB, Pandas, NumPy
Cloud & Systems: AWS, Linux, Git
Supporting / Familiar: FAISS, SQLAlchemy, Streamlit, Experiment Tracking, Model Monitoring

PROFESSIONAL EXPERIENCE

Software Engineer — New Mek Solutions Hyderabad, India
January 2021 – December 2023

- Designed and deployed **ML-backed inference services** using FastAPI and Docker, supporting internal NLP and prediction workflows used by product and analytics teams.
- Built and maintained **5+ REST-based inference endpoints**, handling concurrent requests with stable latency under parallel access.
- Developed **Python and SQL data pipelines** for model training and evaluation, processing datasets ranging from **tens of thousands to low millions of records** depending on use case.
- Implemented **NLP pipelines using Hugging Face Transformers** for document classification, summarization, and information extraction, replacing brittle rule-based logic.
- Reduced average API response latency by **25–35%** through async request handling, query optimization, and improved database access patterns.
- Containerized services and implemented **CI/CD pipelines with GitHub Actions**, reducing manual deployment effort and environment related failures across releases.
- Supported deployed services post-release, investigating **data quality issues, model errors, and performance regressions** in live environments.

PROJECTS

Persistent Memory Layer for LLM Applications GitHub
Python | FastAPI | FAISS | SQLite | SQLAlchemy | LM Studio

- Built a **persistent, task-scoped memory service** for LLM applications, enabling long-term recall while preventing cross-task and cross-user context leakage.
- Implemented **semantic retrieval using FAISS**, indexing **thousands of memory entries per user** and injecting only top-k relevant context based on similarity thresholds.
- Enforced **strict namespace-based isolation** across users and tasks, validated through parallel request and multi-session testing.
- Developed an async FastAPI backend with durable SQLite persistence, ensuring **consistent behavior across restarts and crash scenarios**.
- Integrated **local LLM inference via OpenAI-compatible APIs using LM Studio**, enabling separation of chat and embedding models without external API dependency.
- Reduced average prompt size by **30–40%** by decoupling conversational context from long-term memory recall.

Clinical Communication Memory System GitHub
FastAPI | SQLite | Vector Embeddings | Semantic Search

- Designed a **visit-scoped semantic memory system** for multilingual doctor–patient communication to prevent cross-patient data exposure.
- Enforced **UUID-based visit scoping** at request, service, and repository layers, blocking all reads and writes without explicit visit context.
- Implemented a **fail-closed retrieval strategy**, ensuring embedding or vector search failures returned safe empty results.
- Logged and audited **100% of semantic retrieval events**, enabling traceability for debugging, testing, and compliance review.
- Stress-tested isolation guarantees using **concurrent and adversarial request scenarios**, validating correct behavior under parallel access.

EDUCATION

Saint Louis University St. Louis, MO
Master of Science in Information Systems GPA: 3.90