Krishna Chaitanya Bodepudi

St. Louis, MO (Open to Relocate) | 6468211711 | kcbodepudi@gmail.com | LinkedIn | Portfolio

APPLIED AI / ML ENGINEER

Applied AI / ML Engineer with professional experience building **ML-driven backend systems, AI-enabled services, and production-grade APIs**. Experienced across the full lifecycle of applied ML work, including **model integration, inference services, data pipelines, and system-level reliability**. Comfortable operating at the intersection of **machine learning, backend engineering, and applied AI**, with a strong focus on correctness, scalability, and predictable behavior in real-world environments.

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

ML Engineer Intern — Melotech Remote
January 2024 – Present

- Supported development of **ML-backed inference services** used in internal product and content experimentation workflows.
- Implemented **FastAPI-based APIs** for model inference and embedding-based retrieval with clean request validation and predictable response behavior.
- Built and tested **vector embedding pipelines** to surface relevant contextual information for downstream ML and content-related processes.
- Assisted with **inference-time optimization**, including batching and response-size control, to reduce latency during internal testing and iteration.
- Evaluated model outputs against real media samples, identifying failure cases related to semantic drift, low-confidence retrieval, and malformed inputs.
- Added defensive handling for edge cases in inference and retrieval paths, ensuring systems returned safe fallbacks instead of partial or misleading results.

Software Engineer — New Mek Solutions Hyderabad, India
January 2022 – December 2023

- Designed and deployed **ML-backed inference services** using FastAPI and Docker, supporting internal NLP and analytics workflows.
- Built and maintained **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 from **tens of thousands to low millions of records**.
- Implemented **NLP pipelines using Hugging Face Transformers** for document classification, summarization, and information extraction.
- Reduced average API response latency by **25–35%** through async processing and database query optimization.
- Worked closely with downstream consumers of ML services to refine API contracts, adjust data schemas, and resolve integration issues, improving reliability of model outputs in dependent applications.

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